SPDI & NCBI Variation Data Processing Services

Processing variation data in various formats using the SPDI common data model

https://api.ncbi.nlm.nih.gov/variation/v0

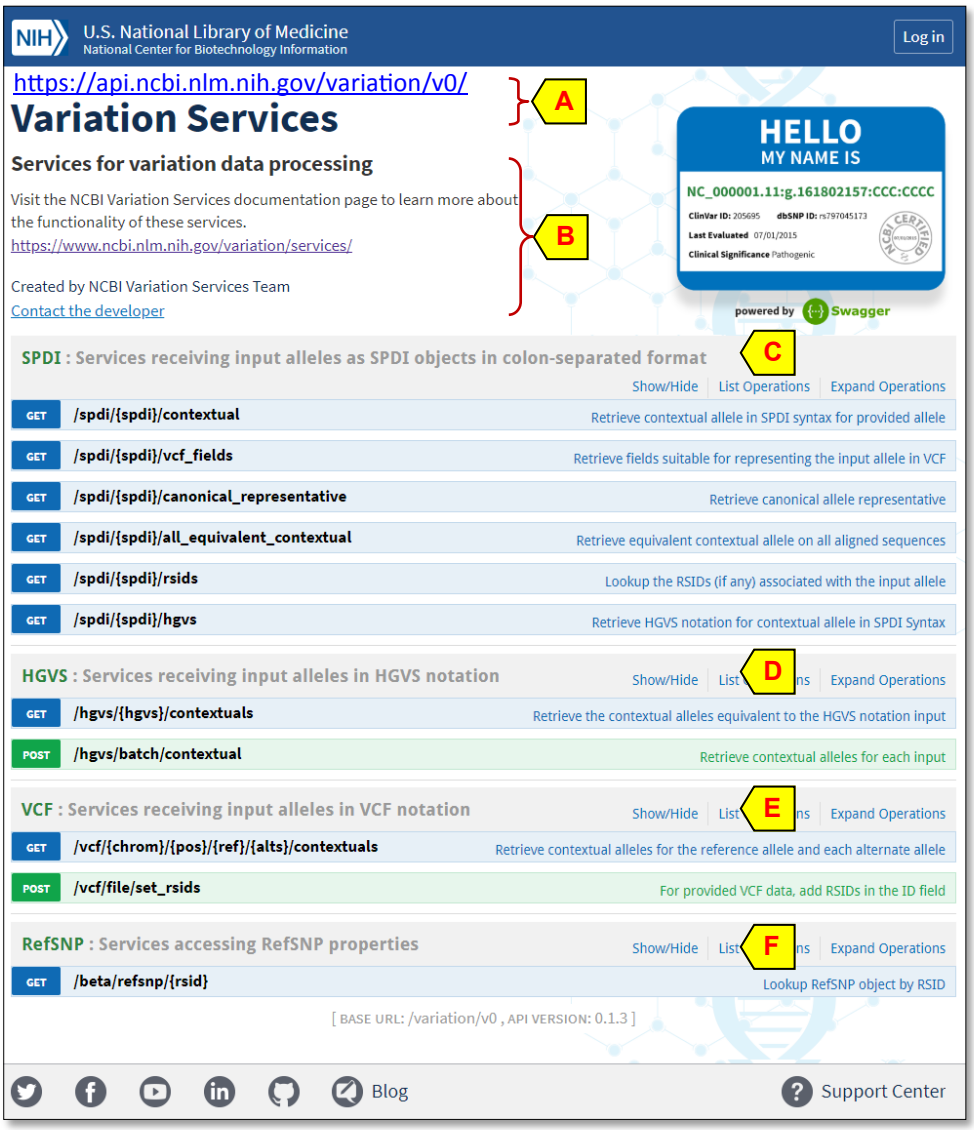
National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Scope and Access

The task of determining whether two genetic variants are the same poses certain challenges. We must navigate competing text file formats for variant representation, different standards for shifting ambiguous alignments, and deal with continually updated reference sequence models on which the variation calls are made. To help address these problems, we are publishing a set of versioned Variation Services for the genomic research community to use to group and compare variant observations. Variation Services use a common data model described as **Sequence Position Deletion Insertion** (SPDI, https://www.ncbi.nlm.nih.gov/variation/notation/). With these services, we are able to

- interconvert between HGVS and VCF formats, with proper left and right-shifting
- determine, if two variants are the same, using the same standard applied by ClinVar and dbSNP
- discover where the variant maps to on the current set of RefSeq sequence models
- retrieve RefSNP identifier by allele using VCF input, and
- retrieve detailed information for a RefSNP record as an JSON object.

For researchers with a one-off analysis need, you can refer to the API Documentation (below) as well as an NCBI blog-post (https://go.usa.gov/xUuJc) for tips on using the API. You can use this set of services as a functional replacement of the recently retired Variation Reporter (https://go.usa.gov/xUJhf). For software developers and workflow engineer, you can incorporate the services into your own data analysis pipelines, which would provide grouping of variants in the same way NCBI variation resources do it. Our initial release (version 0) and every major version thereafter will maintain the backward-compatibility of the object schema.



Accessing the Variation Services

As a start, you can access the Variation Services through its landing page shown to the left (**A**). The top of this page (**B**) provides a summary description and a link to the help document with additional details. The body of the page groups functions available into major categories, each under their own heading:

- The SPDI section (**C**) contains functions that use SPDI as input and returns output in other format such as VCF, HGVS, and RSID.
- The HGVS section (**D**) contains functions that use HGVS as input and return the variant context in SPDI format.
- The VCF section (**E**) contains functions that takes VCF input and returns either the SPDI context or rsids.
- The last RefSNP section (**F**) contains a function that takes an RSID as input and returns a JSON data blob similar to the content used for the newly redesigned RefSNP page, such as https://www.ncbi.nlm.nih.gov/snp/rs268.

Under each heading, available functions appear in separate lines, listing the base URL along with a phrase describing that function.

Contact: info@ncbi.nlm.nih.gov

Accessing the Variation Services (cont.)

The Variation Services landing page displays each available functions by:

- showing the operation mode as a button, with **GET** (**A**) in blue and **POST** in green (POST mode accepts batch input)
- listing its base URL construct (**B**), and
- describing the task the function accomplishes (**C**)

You can click the button to toggle open the display of a function and see example input in the Parameters subsection (**D**), click "Try it out!" button (**E**) to test the service with sample input, and see the actual command line as well as the actual output.

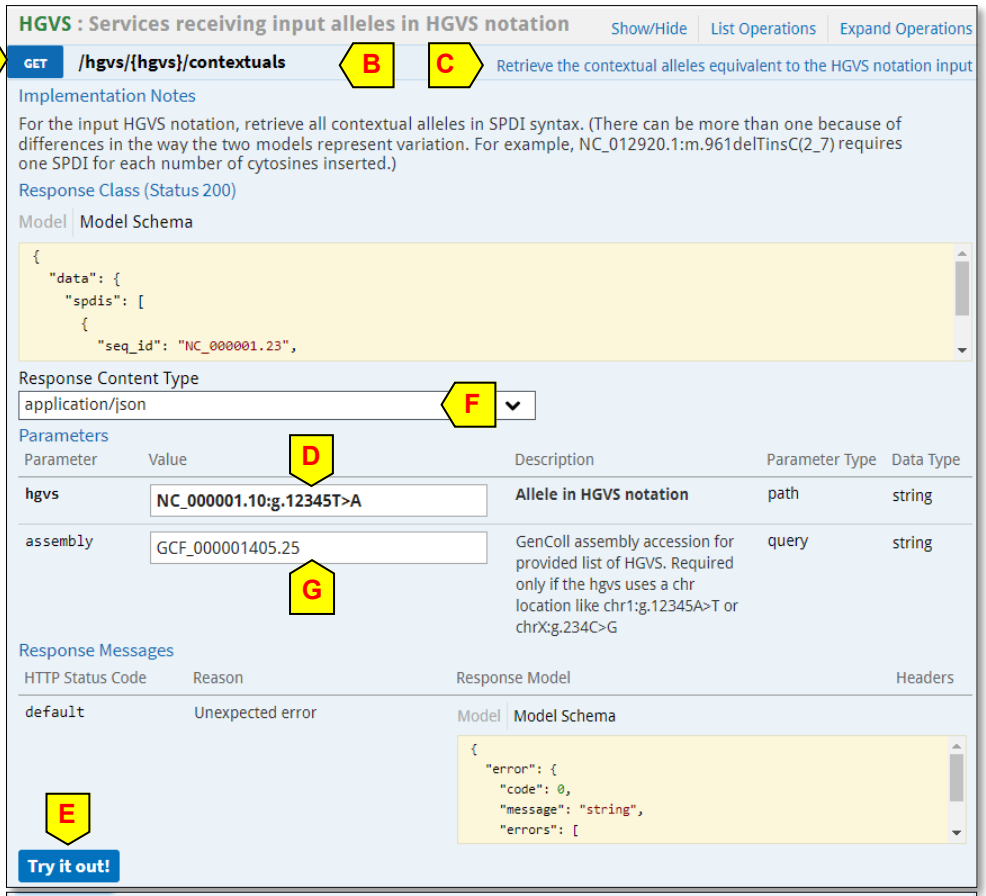
Usage Examples for Functions Taking HGVS and VCF as Input

The third and fourth group of functions provided by the Variation Services take HGVS and VCF as input, respectively.

1. HGVS : Services receiving input alleles in HGVS notation

1a. For single HGVS expression

As the description stated, this service takes variant in HGVS notation (**D**), and retrieves the contextual alleles (in JSON format, **F**). For HGVS without explicit reference to a specific Reference Sequence record (accession. Version), you will need to provide the assembly accession. Version (**G**). If you do not know this for the specific assembly you are using to make your variant calls, you can find it out by searching in the NCBI Assembly database.



Clicking "Try it out!" button (**E**), Variation Services will execute with the provided input and append the result in an added section below. The yellow textboxes (**H**) from the top to bottom provide

- the curl command line with its parameter/value pairs
- the URL for the request, with the unsafe characters escaped
- the services' response in JSON format
- the status code, and
- the response header

You can model the curl command line or the request URL, and incorporate them in your own workflow need.

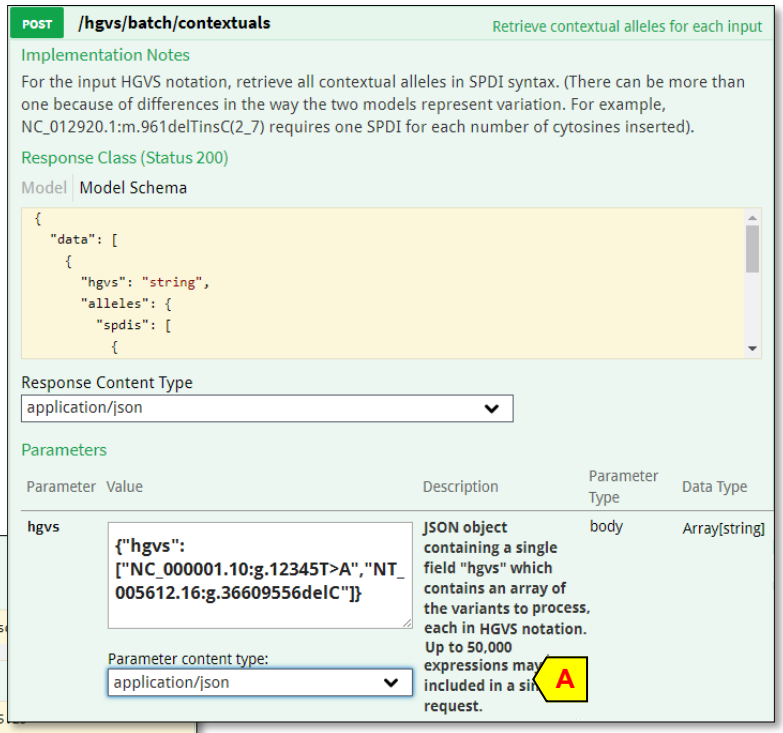
1. HGVS : Services receiving input alleles in HGVS notation (cont.)

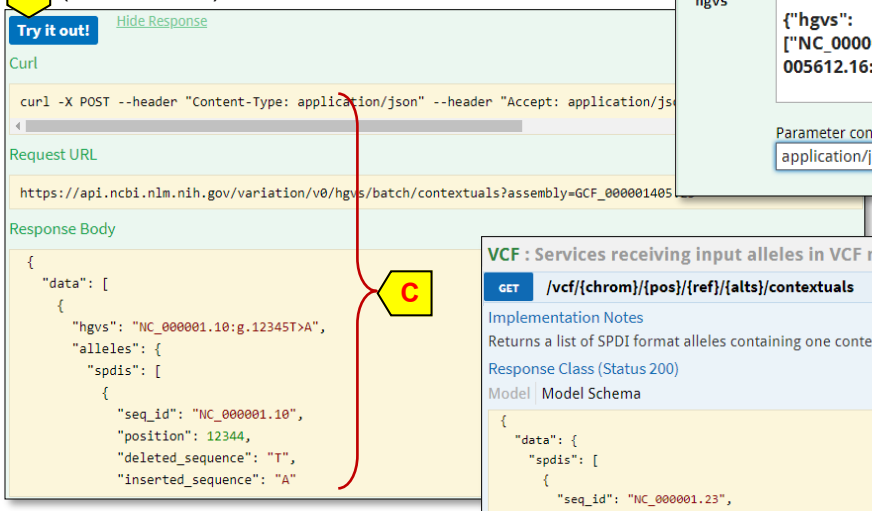
1b. For batch HGVS expressions

This section does the same general task, providing the contextual information for input variant expressed in HGVS format, but use POST mode allows it to take a batch of HGVS input up to 50,000 expressions (**A**). Clicking "Try it out!" button (**B**), Variation Services will execute with the provided input and append the result in an added section below. The yellow textboxes (**C**) from the top to bottom provide

- the curl command line with its parameter/value pairs
- the base-URL for the request (the JSON input is submitted through POST is not included in the URL)
- the services' response in JSON format

The status code and the response header boxes (similar to 1a) are not shown.



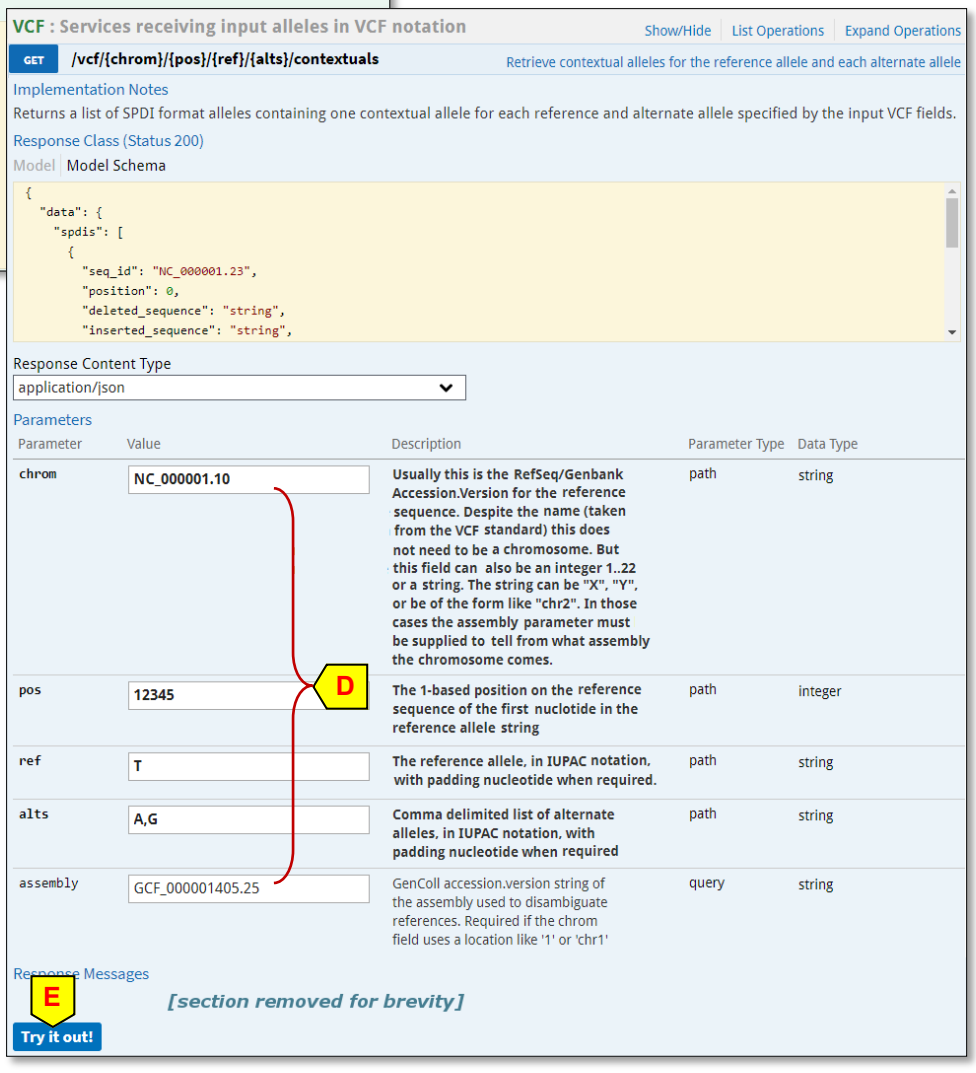


2. VCF : Services receiving input alleles in VCF notation

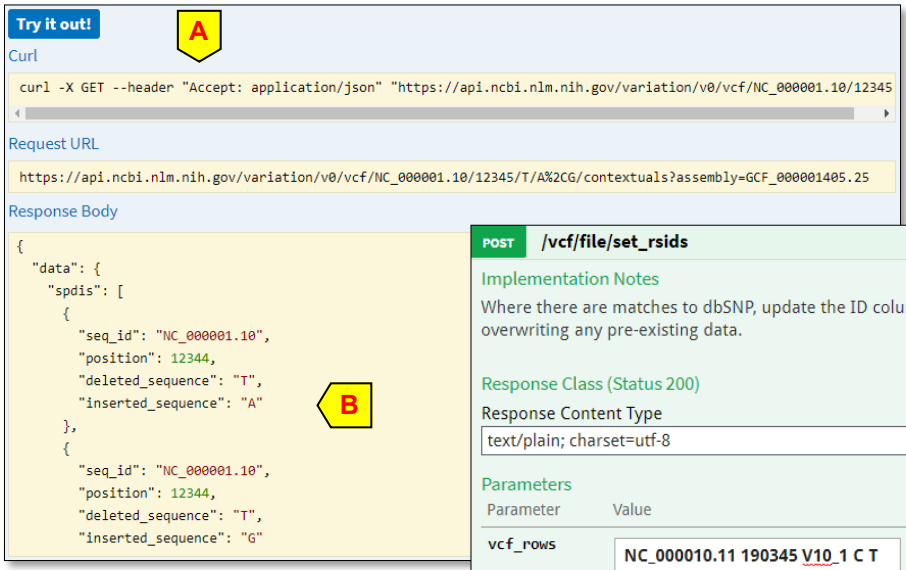
2a. For single variant in VCF format

This service (shown to the right) takes a variant in VCF notation provided through a set of input boxes (**D**). Refer to the Description column for details regarding the value you need to provide to each input fields. For input to chromosome field without explicit reference to a specific Reference Sequence record (accession.version), you will need to provide the assembly accession. You can obtain the information from the NCBI Assembly database.

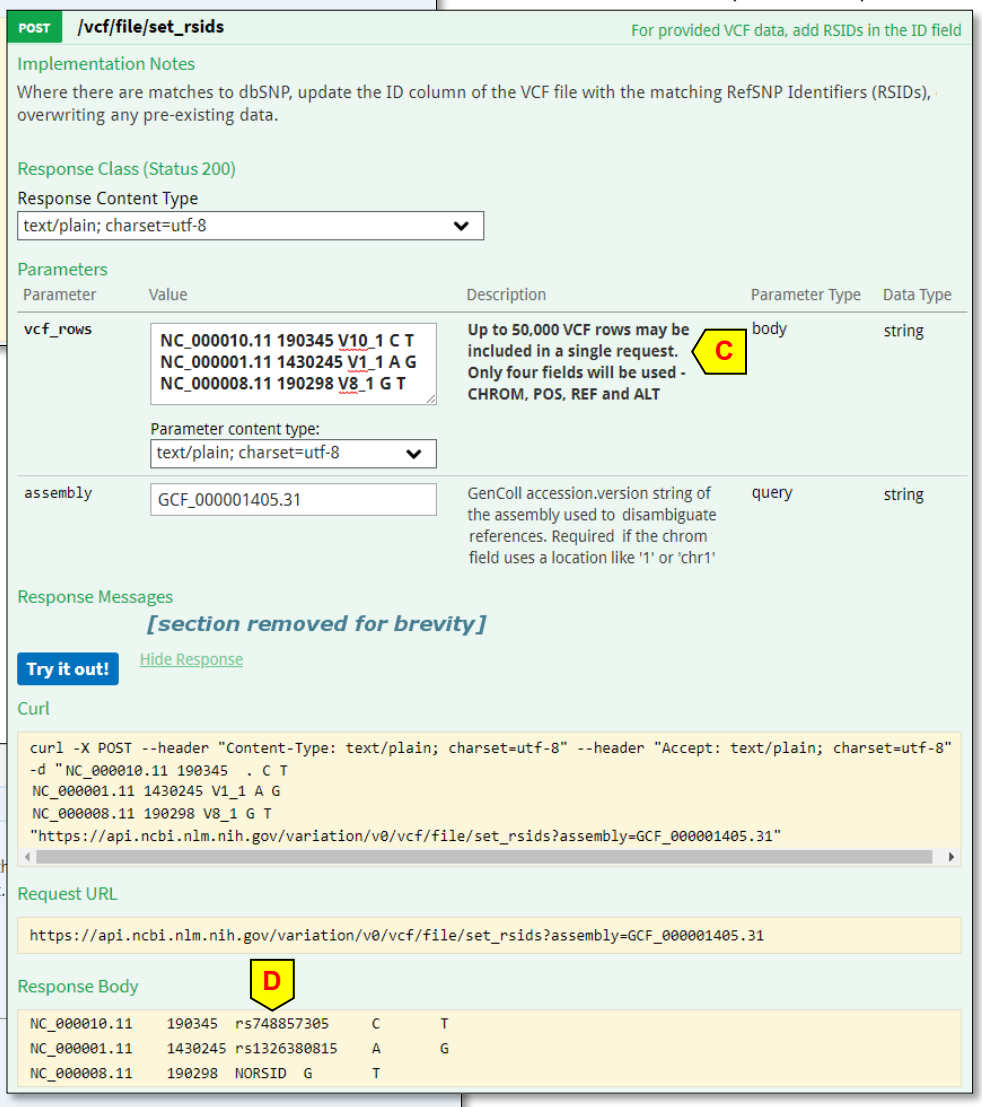
Clicking the "Try it out!" button (**E**), the service will take the input, execute the request, and returns the contextual information in JSON format.



2. VCF : Services receiving input alleles in VCF notation (cont.)

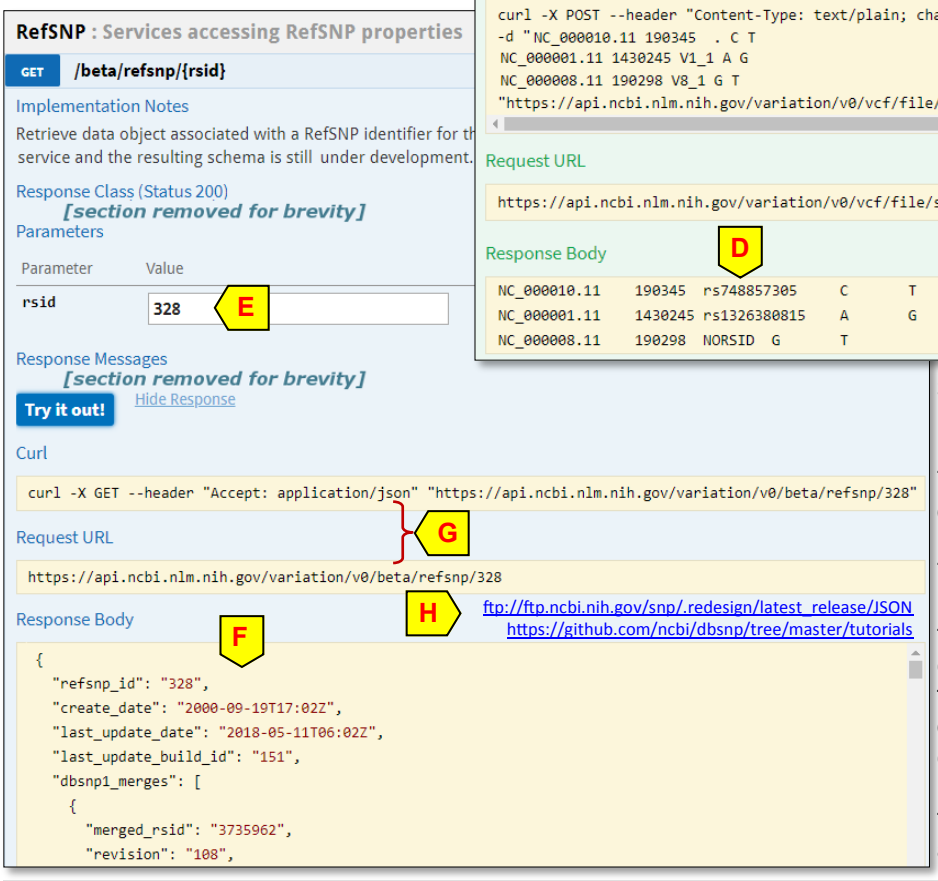


The services appends the result in a new section (**A**) below the button, which contains the curl command line and request URL, followed by actual server response (SPDI allele context in JSON format, **B**). The status code and response header fields are further below (not shown).



2b. For batch VCF-formatted input

The main purpose of this service is to map a set of input variants called in a study to existing entries in dbSNP. You can paste up to 50,00 lines of variants, expressed in 5-column VCF expression, in the input box (**C**). The service overwrites the custom identifiers in the 3rd column with matched RSIDs or NORSID if such mapping fails (**D**).



3. RefSNP : Services accessing RefSNP properties

The service under this section provides direct access to the complete SNP record in JSON format when given a valid RSID. The example (left) uses the RSID 328 as input (**E**) to retrieve the JSON object for that SNP record (**F**). You can readily incorporate the function in your own workflow by modeling after the command line or request URL examples (**G**). For the complete SNP dataset in JSON format, use the FTP site instead. Tutorials on how to process the JSON dataset is also available online (**H**).